A shield with a mermaid and lion

Description automatically generated

**Unsupervised Anomaly Detection in Brain MRI Scans Using 3D Convolutional Autoencoders**

**Stanley Allen**

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

2023-2024

Abstract

Neurodegenerative diseases, such as Alzheimer's disease and mild cognitive impairment (MCI), pose significant global health challenges due to their progressive nature and lack of curative treatments. Early detection is crucial for effective intervention, yet identifying subtle brain anomalies at early stages remains a challenge in medical imaging. This study investigates the use of a 3D Convolutional Autoencoder (CAE) for unsupervised anomaly detection in brain MRI scans, focusing on improving detection accuracy by combining Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM) as evaluation metrics.

The CAE was trained exclusively on healthy brain images to learn the typical features of a healthy brain and was evaluated across three major datasets: the Human Connectome Project (HCP), Alzheimer’s Disease Neuroimaging Initiative (ADNI), and Cambridge Centre for Ageing and Neuroscience (CamCAN). The model's performance was assessed on 223 images from each dataset, highlighting its ability to generalize across different datasets and age groups.

Results showed that while the CAE effectively reconstructed healthy brain structures with high accuracy, significant reconstruction errors were observed at the boundaries between brain matter and non-brain regions, particularly in older adults. These errors underscored the limitations of relying solely on MSE and SSIM for anomaly detection, indicating a need for more advanced metrics or supplementary techniques. The study concludes with recommendations for enhancing the model's robustness, potentially integrating it with other neural network architectures like ResNet for improved clinical applicability.

Acknowledgements

I would like to express my deepest gratitude to all those who contributed to the completion of this dissertation.

First and foremost, I extend my sincere thanks to my supervisor, Melanie Jouaiti, for her unwavering support, guidance, and encouragement throughout this project. Her insightful feedback and expertise were invaluable in shaping the direction and quality of this research.

I am also grateful to the faculty and staff at University of Birmingham for providing the resources and academic environment necessary to complete this work. The knowledge and skills I have acquired during my time here have been instrumental in conducting this research.

I would like to acknowledge the teams behind the datasets used in this research: the Human Connectome Project (HCP), the Alzheimer’s Disease Neuroimaging Initiative (ADNI), and the Cambridge Centre for Ageing and Neuroscience (CamCAN). Their comprehensive and well-maintained datasets provided the foundation for the analysis conducted in this study.

I also wish to thank my colleagues and fellow students, whose discussions and feedback provided additional perspectives and motivation during this work.

A special thanks to my family and friends for their constant encouragement, understanding, and patience throughout this journey. Their support has been a source of strength for me.

Abbreviations

**AD** - Alzheimer’s Disease  
**ADNI** - Alzheimer’s Disease Neuroimaging Initiative  
**AUC** - Area Under the Curve  
**CAE** - Convolutional Autoencoder  
**CamCAN** - Cambridge Centre for Ageing and Neuroscience  
**CLAHE** - Contrast Limited Adaptive Histogram Equalization  
**CNN** - Convolutional Neural Network  
**HCP** - Human Connectome Project  
**MCI** - Mild Cognitive Impairment  
**MSE** - Mean Squared Error  
**MRI** - Magnetic Resonance Imaging  
**ReLU** - Rectified Linear Unit  
**ResNet** - Residual Network  
**SSIM** - Structural Similarity Index Measure

List of Figures

Figure 1.

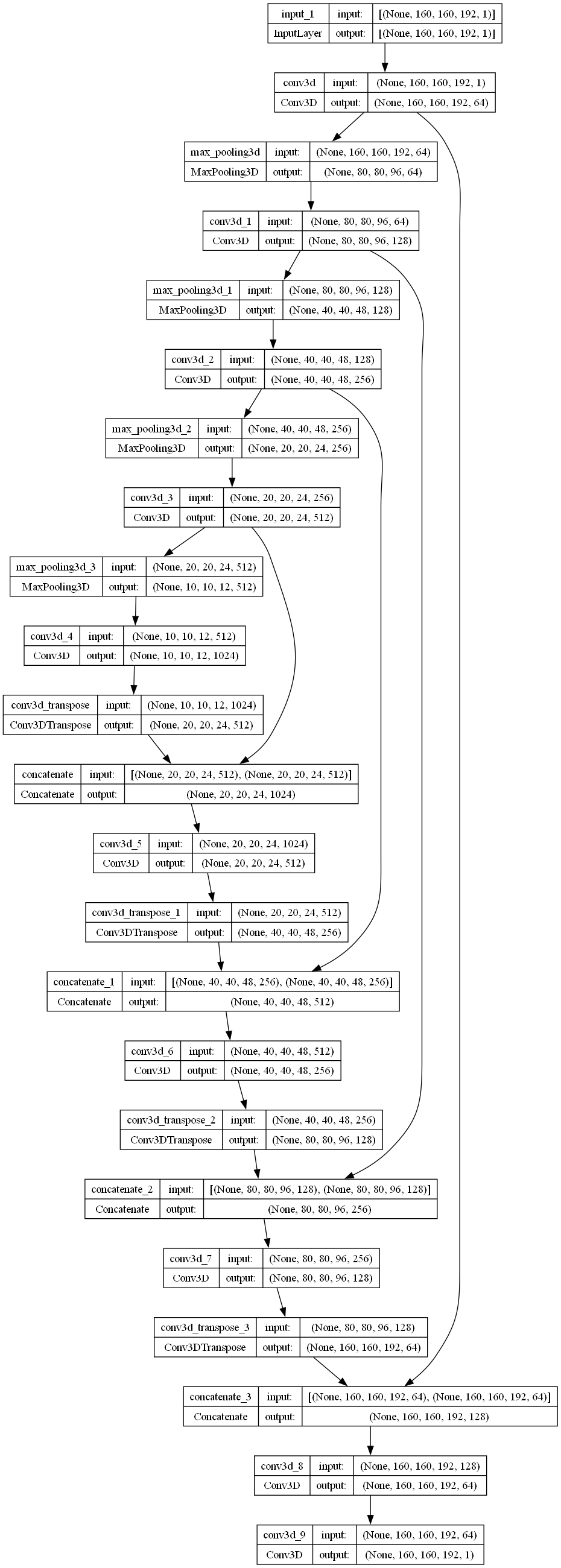


Figure 2.

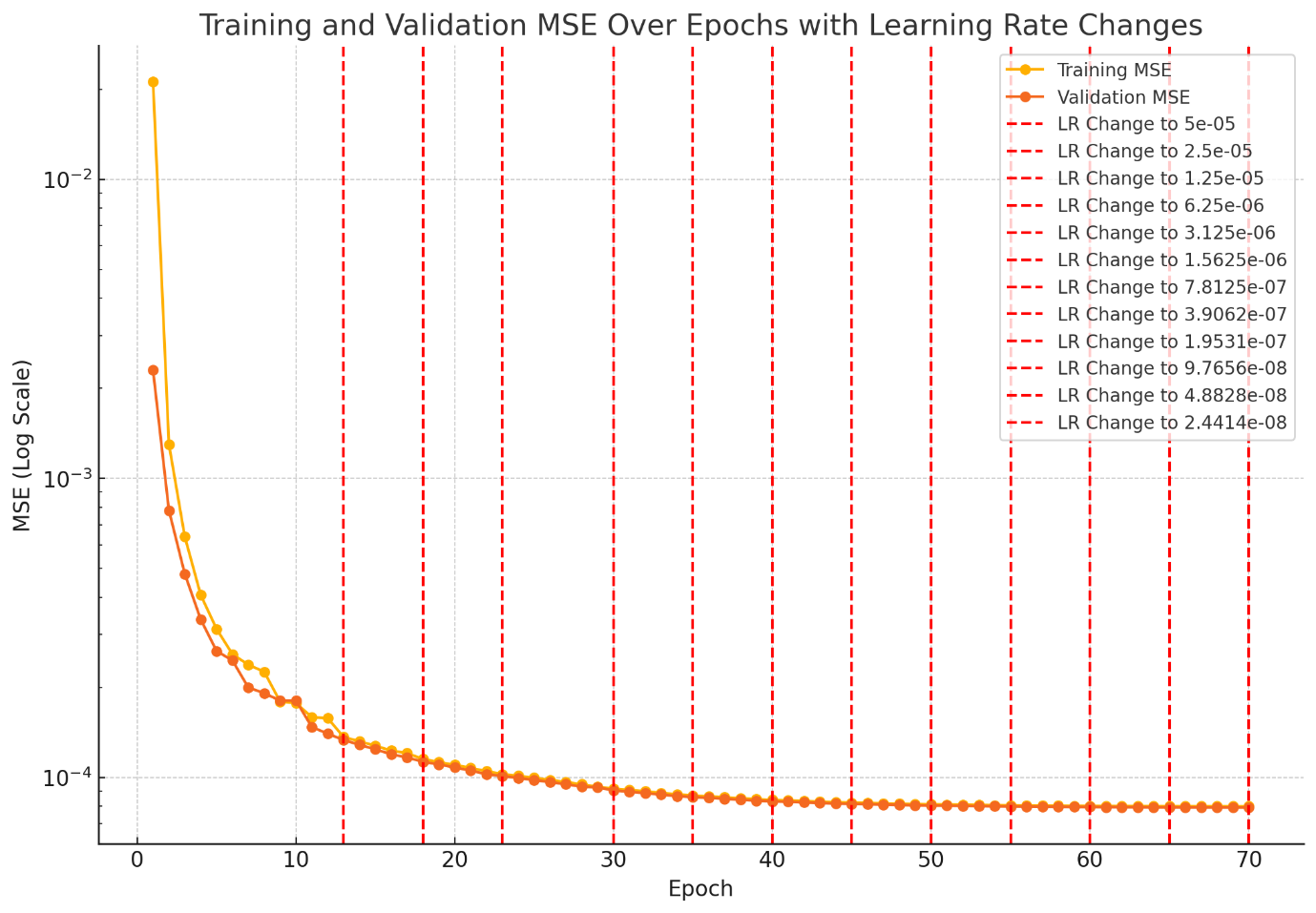


Figure 3.

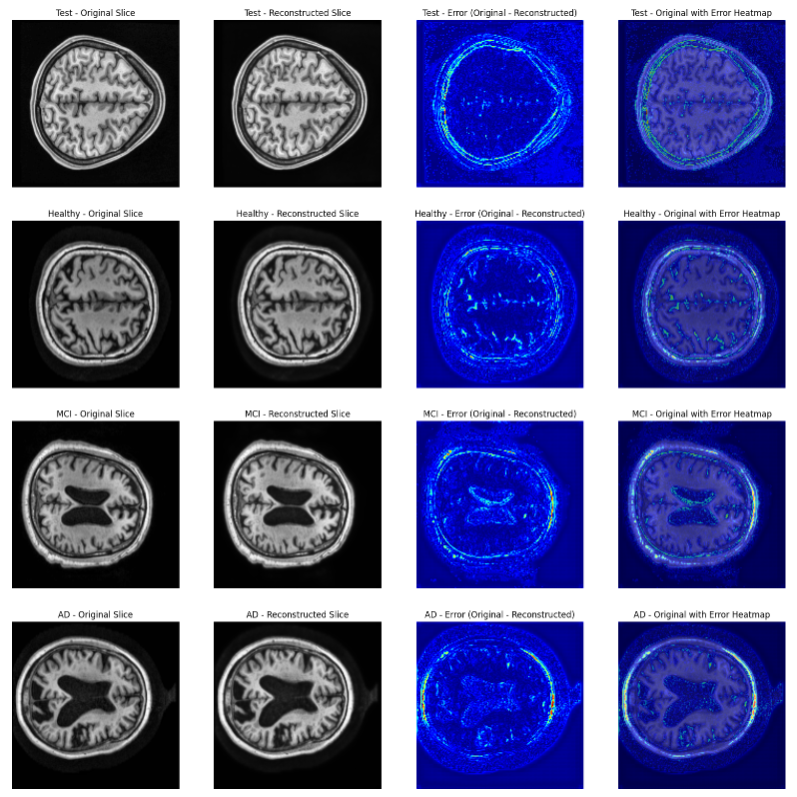


Figure 4.

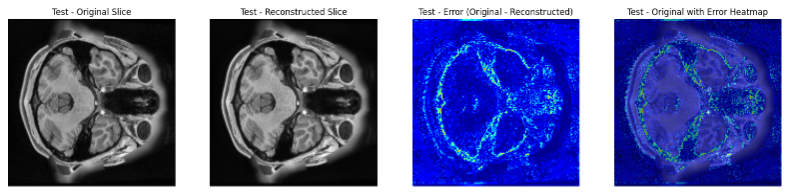


Figure 5.

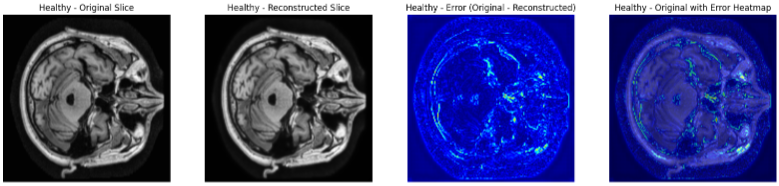


Figure 6.

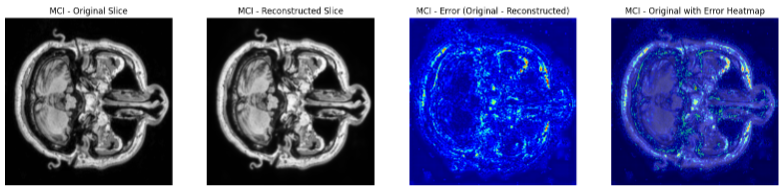


Figure 7.

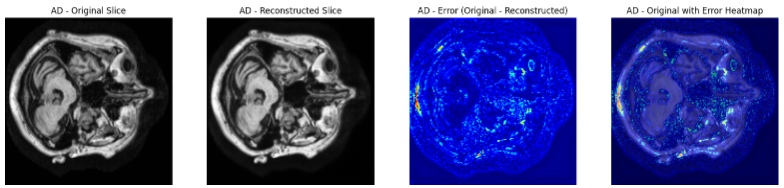


Figure 8 (under 40).

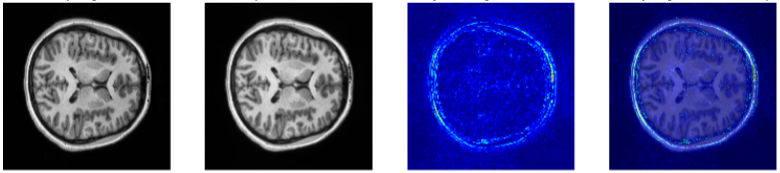


Figure 9 (over 60).

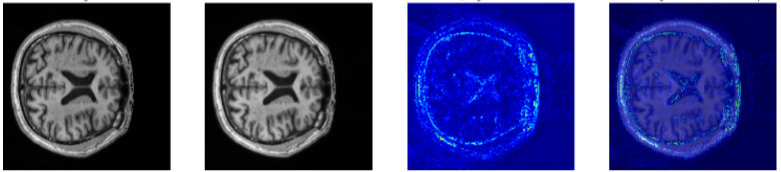


Figure 10.

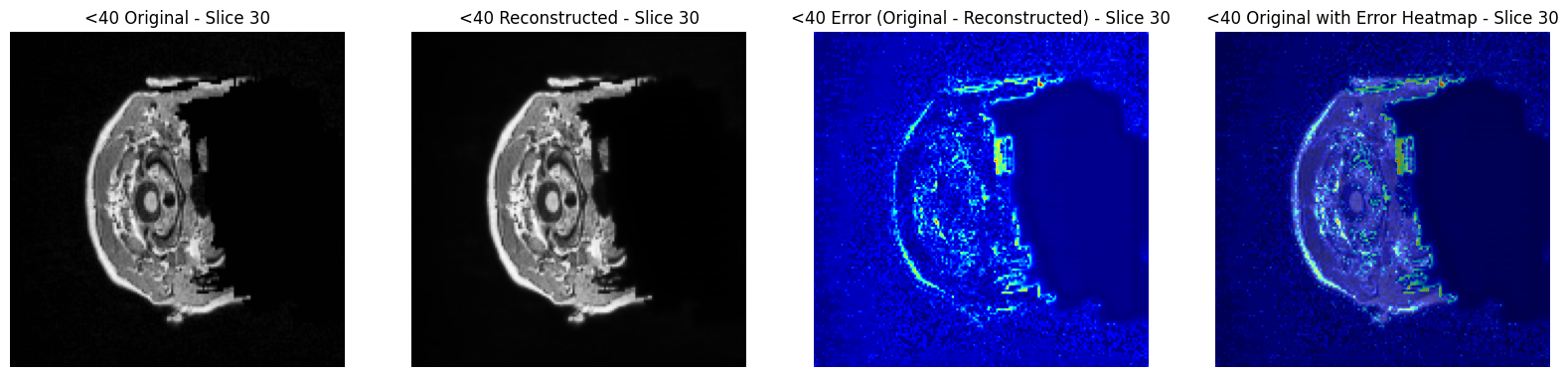
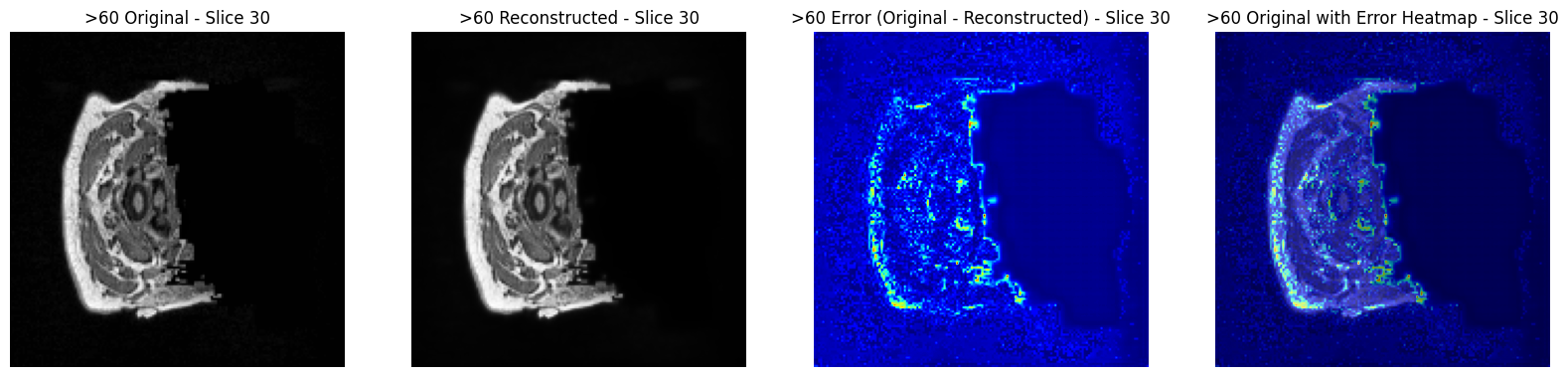


Figure 11.



Contents

[1. Introduction 13](#_Toc176147921)

[1.1 Background and Motivation 13](#_Toc176147922)

[1.2 Problem Statement 15](#_Toc176147923)

[1.3 Research Objectives 17](#_Toc176147924)

[1.4 Significance of the Study 18](#_Toc176147925)

[1.5 Ethical Considerations in Using MRI Data 19](#_Toc176147926)

[1.6 Dissertation Structure 19](#_Toc176147927)

[2. Literature Review 21](#_Toc176147928)

[2.1 Brain Atrophy and Neurodegenerative Diseases 21](#_Toc176147929)

[2.2 Anomaly Detection in Medical Imaging 22](#_Toc176147930)

[2.3 Autoencoders in Medical Imaging 23](#_Toc176147931)

[2.4 Use of SSIM and MSE in Evaluating Autoencoder Performance 25](#_Toc176147932)

[2.5 Data Augmentation and Preprocessing Techniques 26](#_Toc176147933)

[2.6 Baseline Models and Comparison with Previous Research 26](#_Toc176147934)

[2.7 Applications and Case Studies in Medical Imaging 27](#_Toc176147935)

[2.8 Conclusion 29](#_Toc176147936)

[3. Methodology 30](#_Toc176147937)

[3.1 Introduction 30](#_Toc176147938)

[3.2 MRI Technology Overview 30](#_Toc176147939)

[3.3 Data Acquisition and Quality Challenges 31](#_Toc176147940)

[3.4 Data Preprocessing 33](#_Toc176147941)

[3.4.1 Contrast Enhancement 33](#_Toc176147942)

[3.4.2 Intensity Normalization and Resizing 33](#_Toc176147943)

[3.4.3 Image Reorientation 34](#_Toc176147944)

[3.5 Model Architecture 34](#_Toc176147945)

[3.5.1 Encoder 34](#_Toc176147946)

[3.5.2 Bottleneck 34](#_Toc176147947)

[3.5.3 Decoder 35](#_Toc176147948)

[3.5.4 Model Architecture Diagram 35](#_Toc176147949)

[3.6 Training Procedure 35](#_Toc176147950)

[3.6.1 Training Configuration 36](#_Toc176147951)

[3.6.2 ReduceLROnPlateau and Early Stopping 37](#_Toc176147952)

[3.7 Evaluation Metrics 37](#_Toc176147953)

[3.7.1 Mean Squared Error (MSE) 38](#_Toc176147954)

[3.7.2 Structural Similarity Index Measure (SSIM) 39](#_Toc176147955)

[3.8 Visualization Techniques 39](#_Toc176147956)

[3.8.1 Image Reconstruction Visualization 39](#_Toc176147957)

[3.8.2 Heatmaps 40](#_Toc176147958)

[3.9 Baseline Models and Comparative Analysis 40](#_Toc176147959)

[3.9.1 Selection of Baseline Models 40](#_Toc176147960)

[3.9.2 Performance Comparison 41](#_Toc176147961)

[3.9.3 Limitations of Baseline Models 42](#_Toc176147962)

[3.10 Model Evaluation and Results 42](#_Toc176147963)

[3.10.1 Test Data Evaluation 43](#_Toc176147964)

[3.10.2 External Dataset Evaluation 43](#_Toc176147965)

[4. Results 45](#_Toc176147966)

[4.1 Training Curve Analysis 45](#_Toc176147967)

[4.1.1 Training Logs 45](#_Toc176147968)

[4.1.2 Training Curve Visualization 45](#_Toc176147969)

[4.2 Performance on Human Connectome Dataset 46](#_Toc176147970)

[4.3 Anomaly Detection in ADNI Dataset 48](#_Toc176147971)

[4.4 Case Study on CamCAN Dataset 49](#_Toc176147972)

[4.5 Comparison with Baseline Models 51](#_Toc176147973)

[4.6 Visualization of Reconstruction Errors 52](#_Toc176147974)

[4.7 Limitations in Error Masking 52](#_Toc176147975)

[5. Discussion 53](#_Toc176147976)

[5.1 Interpretation of Results 53](#_Toc176147977)

[5.2 Comparison with Existing Methods 54](#_Toc176147978)

[5.3 Strengths and Limitations 55](#_Toc176147979)

[5.4 Implications for Clinical Practice 57](#_Toc176147980)

[6. Conclusion 58](#_Toc176147981)

[6.1 Summary of Findings 58](#_Toc176147982)

[6.2 Contributions to the Field 59](#_Toc176147983)

[6.3 Future Work 59](#_Toc176147984)

[References 62](#_Toc176147985)

[Appendix 66](#_Toc176147986)

[GitLab Repository Address 66](#_Toc176147987)

[Contents of the GitLab Repository 66](#_Toc176147988)

# 1. Introduction

## 1.1 Background and Motivation

Neurodegenerative diseases such as Alzheimer's disease, mild cognitive impairment (MCI), and Parkinson's disease represent a significant global health burden due to their progressive nature and lack of curative treatments. These conditions are characterized by the gradual degeneration of neurons, leading to cognitive and motor decline. Brain atrophy, a reduction in brain volume, is a key indicator of these diseases and can be detected using Magnetic Resonance Imaging (MRI) (Baur et al., 2019).

MRI is a non-invasive imaging technique widely used in clinical settings to diagnose and monitor the progression of neurodegenerative diseases (McRobbie et al., 2007). It provides high-resolution images of the brain's structure, allowing clinicians to assess the extent of brain atrophy and identify abnormal structures. However, the vast amount of data generated by MRI scans poses challenges in manual analysis, which is time-consuming and prone to error. Additionally, subtle structural changes in the brain, particularly in the early stages of neurodegenerative diseases, may be overlooked using traditional analysis methods (Ravishankar et al., 2017).

To address these challenges, machine learning techniques, particularly autoencoders—a type of unsupervised neural network—have gained traction for detecting anomalies in medical images. Autoencoders compress input data into a lower-dimensional representation and then reconstruct it. When trained on healthy brain images, they learn the typical features of a healthy brain. If an image deviates from these features, the reconstruction process yields a higher error, flagging potential anomalies indicative of disease (Larsen et al., 2015).

Reconstruction error is typically measured using Mean Squared Error (MSE), which quantifies the difference between the original and reconstructed images. Although MSE is a straightforward and commonly used metric, it has limitations in the context of medical imaging. MSE treats all pixel differences equally, without considering their clinical significance. This can result in the failure to detect subtle but important anomalies or the incorrect classification of normal variations as abnormalities (Bergmann et al., 2019). To overcome these limitations, the Structural Similarity Index Measure (SSIM) is employed alongside MSE, as SSIM evaluates images based on structural and perceptual differences, providing a more comprehensive assessment (Wang et al., 2004).

In this study, SSIM was combined with MSE to enhance the accuracy and reliability of anomaly detection in brain MRI scans during the evaluation phase. While MSE was the only metric used during the training phase to monitor the model's performance, SSIM was introduced alongside MSE during the evaluation to provide a more nuanced assessment of the reconstructed images. The model was trained exclusively on a healthy dataset using the BlueBEAR supercomputer, with the goal of ensuring that it accurately learned the typical features of a healthy brain.

The model was evaluated on both the test dataset (a subset of the data from which the training dataset was derived) and external datasets, including the Alzheimer’s Disease Neuroimaging Initiative (ADNI). Although the model demonstrated strong performance in terms of structural similarity, the accuracy required for anomaly classification (distinguishing between healthy and abnormal brain scans) proved challenging. This difficulty arose primarily due to the presence of too many variables and reconstruction errors in the non-brain matter, which led to unreliable results. Despite these challenges, the model generally performed well in terms of generalization across different datasets, underscoring the importance of using diverse data to further enhance the model’s robustness and accuracy.

## 1.2 Problem Statement

Detecting brain anomalies, especially in the early stages of neurodegenerative diseases, remains a significant challenge in medical imaging. Early detection is crucial for effective intervention, yet these early anomalies are often the most difficult to identify and may be overlooked due to their subtlety. Traditional methods for anomaly detection typically rely on supervised learning, which necessitates labelled datasets that contain examples of both healthy and abnormal cases. However, obtaining sufficiently labelled datasets is challenging, particularly for subtle anomalies that characterize early neurodegenerative diseases (Litjens et al., 2017). Furthermore, the complexity of brain structures and the variability in MRI scan quality can exacerbate the difficulty of accurately classifying these anomalies. As suggested by Lundervold and Lundervold (2019), the classification of MRI scans is often complicated by factors such as noise, artifacts, and the intrinsic variability in human anatomy, which can lead to inconsistencies in diagnosis and hinder the effectiveness of traditional supervised learning methods.

This research addresses these challenges by developing an unsupervised learning approach using a 3D convolutional autoencoder trained exclusively on healthy brain MRI scans. The autoencoder is designed to reconstruct the typical features of a healthy brain, and deviations from these features result in increased reconstruction error, signalling potential anomalies. This approach is advantageous in medical imaging applications where labelled datasets are scarce, as it does not require labelled data for training (Schlegl et al., 2017).

The effectiveness of this approach hinges on the metrics used to evaluate the reconstruction error. While MSE is the standard metric, it has limitations in distinguishing between normal and abnormal brain scans. To address these limitations, this study also employs SSIM, which provides a more nuanced assessment by considering structural and perceptual differences. The combination of MSE and SSIM is expected to improve the accuracy of anomaly detection.

The performance of the autoencoder will be evaluated by comparing it to findings from previous studies in the field. This comparison is crucial for validating the proposed approach and ensuring it aligns with or improves upon existing methods. The results could have significant implications for the early detection of neurodegenerative diseases, potentially leading to earlier and more accurate diagnoses.

## 1.3 Research Objectives

The primary objectives of this research are as follows:

1. **Develop a 3D Convolutional Autoencoder:** Design and implement a 3D convolutional autoencoder tailored for detecting anomalies in brain MRI scans. The model will be trained exclusively on healthy brain images to learn the typical patterns and features associated with a healthy brain. This will enable the model to detect deviations from these patterns as potential anomalies when presented with new data.
2. **Evaluate Metrics for Anomaly Detection:** Critically assess the use of MSE and SSIM as metrics for anomaly detection in medical imaging. While MSE is traditionally used to evaluate reconstruction error, it may not be sufficient for detecting subtle anomalies in complex brain images. SSIM, which considers structural and perceptual differences, will be used alongside MSE to provide a more comprehensive assessment of reconstruction quality.
3. **Compare Against Findings from Previous Studies:** Evaluate the performance of the autoencoder by comparing it to results reported in previous studies. This comparison will help validate the effectiveness of the proposed approach and ensure it meets or exceeds the performance of existing methods. The findings will contribute to the broader understanding of how unsupervised learning techniques can be applied to medical imaging.
4. **Conduct a Case Study Using the CamCAN Dataset:** To assess the generalizability of the model, a case study will be conducted using the CamCAN dataset. This dataset offers a diverse set of brain MRI scans, which will be used to evaluate how well the model performs on data that differs from the training set. The case study will provide insights into the model’s robustness and potential for real-world application.

## 1.4 Significance of the Study

This study aims to make significant contributions to the field of medical imaging and the early detection of neurodegenerative diseases. By developing a 3D convolutional autoencoder and evaluating its performance using both MSE and SSIM, this research seeks to address the limitations of current anomaly detection methods. The integration of SSIM alongside MSE is expected to improve the accuracy and reliability of anomaly detection, making it easier to identify subtle changes in brain structure that may indicate the early stages of disease.

The comparison with previous studies will serve as a benchmark for assessing the effectiveness of the proposed approach. If successful, this method could be integrated into clinical practice, assisting clinicians in identifying early-stage anomalies that might otherwise go undetected. Early detection is crucial for managing neurodegenerative diseases, as it allows for earlier intervention and potentially better patient outcomes.

Furthermore, the case study using the CamCAN dataset will demonstrate the generalizability of the model, highlighting its potential for application across different datasets and clinical settings. This is particularly important for ensuring that the model can be used in various contexts, not just with the specific data on which it was trained.

In summary, this study has the potential to advance the field of medical imaging by providing a more effective method for detecting anomalies in brain MRI scans. The findings could lead to earlier and more accurate diagnoses of neurodegenerative diseases, ultimately improving patient outcomes.

## 1.5 Ethical Considerations in Using MRI Data

In this research, strict ethical guidelines were adhered to concerning the use of MRI data. A Data Use Certificate (DUC) was signed by my supervisor and university officials to ensure the ethical handling and access to the data. Anonymization measures, such as masking identifiable features in the MRI scans, were implemented to protect participants' identities in the CamCAN dataset. This ensured that the study complied with ethical standards, safeguarding the privacy and confidentiality of the subjects involved.

## 1.6 Dissertation Structure

This dissertation is organized into several chapters, each addressing a specific aspect of the research. Following this introduction, Chapter 2 provides a comprehensive literature review on brain atrophy, anomaly detection in medical imaging, and the use of autoencoders for unsupervised learning. This chapter will also discuss the limitations of MSE and the advantages of using SSIM in medical imaging.

Chapter 3 covers the methodology, including the development of the 3D convolutional autoencoder, the datasets used, and the evaluation metrics. This chapter will also outline the procedures for training the model and assessing its performance.

Chapter 4 presents the results of the study, including a detailed comparison of the autoencoder’s performance against findings from previous studies. This chapter will also include the results of the case study using the CamCAN dataset, providing insights into the model’s generalizability.

Chapter 5 discusses the implications of the findings, including their potential impact on clinical practice and future research directions. This chapter will also address the limitations of the study and suggest areas for further investigation.

Finally, Chapter 6 concludes the dissertation by summarizing the key findings and their significance. The dissertation will also include references to all sources cited throughout the work, ensuring that the research is grounded in the existing body of knowledge.

# 2. Literature Review

## 2.1 Brain Atrophy and Neurodegenerative Diseases

Neurodegenerative diseases, including Alzheimer’s disease, mild cognitive impairment (MCI), and Parkinson’s disease, are characterized by the progressive degeneration of neurons, leading to the loss of cognitive and motor functions. A key indicator of these conditions is brain atrophy, which refers to the reduction in brain volume due to the loss of neurons and the connections between them. Brain atrophy is a critical biomarker that can be detected using advanced imaging techniques like Magnetic Resonance Imaging (MRI). MRI provides detailed images of the brain's structure, enabling clinicians to observe and measure these changes non-invasively (McRobbie et al., 2007).

Alzheimer’s disease, the most common form of dementia, is associated with widespread atrophy, particularly in the hippocampus and cortex. Early detection of atrophy in these regions is vital for diagnosis and intervention, as it correlates strongly with the progression of cognitive decline. However, the early stages of Alzheimer's disease often present with subtle atrophic changes, which are challenging to detect with conventional imaging techniques (Jack et al., 2010). MCI, considered a precursor to Alzheimer’s, also involves region-specific brain atrophy, although these changes are typically less severe than those observed in Alzheimer's disease. The progression from MCI to Alzheimer’s underscores the need for early and accurate detection methods (Petersen et al., 2009).

Parkinson’s disease, although primarily known for its motor symptoms, also involves brain atrophy, particularly in the substantia nigra and the basal ganglia. These structural changes can be visualized using MRI, providing insights into the disease’s progression (Zhang et al., 2015). The early detection of brain atrophy in Parkinson's and other neurodegenerative diseases is crucial, as it allows for timely therapeutic interventions that may slow the disease's progression.

Traditional methods for detecting brain atrophy, such as manual inspection and volumetric measurements, are often insufficient, particularly in the early stages of disease when changes are subtle. These methods are also time-consuming and prone to human error, leading to inconsistencies in diagnosis (Ravishankar et al., 2017). Given these challenges, there is a pressing need for more advanced, automated techniques that can accurately detect early-stage atrophy, enabling earlier diagnosis and intervention (Sperling et al., 2011).

## 2.2 Anomaly Detection in Medical Imaging

Anomaly detection in medical imaging is a critical task, particularly for identifying diseases at an early stage when treatment can be most effective. Traditional anomaly detection methods typically rely on supervised learning, where models are trained on labelled datasets containing examples of both normal and abnormal cases. In neuroimaging, however, obtaining labelled datasets is challenging due to the complexity of brain structures and the subtlety of early-stage anomalies. The variability in brain anatomy across individuals further complicates the creation of robust labelled datasets for training models (Litjens et al., 2017).

To address these challenges, unsupervised learning methods, which do not require labelled data, have gained popularity. Among these methods, autoencoders have emerged as a promising tool for anomaly detection in medical imaging. Autoencoders are a type of neural network designed to learn a compressed representation of input data and then reconstruct it. When trained on images of healthy brains, an autoencoder learns to recognize the typical features of a healthy brain. When it encounters an image that deviates from this learned norm, the reconstruction error increases, signalling the presence of a potential anomaly (Schlegl et al., 2017).

The use of autoencoders for anomaly detection in neuroimaging is particularly advantageous because it allows for the identification of abnormalities without the need for extensive labelled datasets. This capability is crucial for early-stage diseases, where labelled examples of anomalies may be rare or difficult to obtain. By training on healthy data, the autoencoder can detect subtle deviations that might indicate the early onset of neurodegenerative diseases (Baur et al., 2019).

## 2.3 Autoencoders in Medical Imaging

Autoencoders have been applied to various tasks in medical imaging, including image denoising, segmentation, and most notably, anomaly detection. The typical architecture of an autoencoder consists of two main components: an encoder, which compresses the input data into a lower-dimensional representation, and a decoder, which reconstructs the original data from this compressed representation. The difference between the original and reconstructed images is quantified using a loss function, such as Mean Squared Error (MSE), which measures the reconstruction error (Baur et al., 2019).

In the context of neuroimaging, autoencoders are used to detect anomalies by training the model on a dataset of healthy brain MRI scans. After training, the autoencoder can be applied to new images, where it attempts to reconstruct them based on the patterns it has learned. If the image represents a healthy brain, the reconstruction error will typically be low. However, if the image contains abnormalities—such as those found in patients with Alzheimer’s disease or other neurodegenerative conditions—the reconstruction error will be higher, indicating a potential anomaly (Larsen et al., 2015).

Despite their effectiveness, the use of autoencoders in medical imaging is not without challenges. One of the primary limitations is the reliance on MSE as the metric for evaluating reconstruction error. MSE is a pixel-wise metric that treats all pixel differences equally, regardless of their structural or clinical significance. This approach can lead to insensitivity to subtle but important anomalies, which are crucial for early diagnosis (Bergmann et al., 2019). To address these limitations, researchers have begun to explore alternative metrics, such as the Structural Similarity Index Measure (SSIM), which evaluates images based on their structural content, providing a more accurate assessment of image quality in medical imaging applications (Wang et al., 2004).

## 2.4 Use of SSIM and MSE in Evaluating Autoencoder Performance

The Structural Similarity Index Measure (SSIM) is a perceptual metric that assesses image quality by comparing structural similarities between an original image and its reconstruction. Unlike MSE, which only accounts for pixel-by-pixel differences, SSIM evaluates three key aspects of the image: luminance, contrast, and structure. This makes SSIM more aligned with human visual perception, which is particularly important in medical imaging, where the preservation of structural details is crucial (Wang et al., 2004).

In medical imaging, combining SSIM with MSE allows for a more robust evaluation of reconstructed images. While MSE can detect gross errors in reconstruction, SSIM captures finer details related to image structure, which are essential in identifying early-stage neurodegenerative changes. This combined approach has proven effective in improving the accuracy of anomaly detection in brain MRI scans. By using both metrics, researchers can ensure that the autoencoder not only minimizes pixel-wise errors but also preserves the structural integrity of the image, leading to more reliable and clinically relevant results (Bergmann et al., 2019).

## 2.5 Data Augmentation and Preprocessing Techniques

Effective preprocessing and data augmentation are critical for enhancing the performance of autoencoders in medical imaging. MRI data inherently varies due to differences in scanner settings, patient positioning, and noise. To standardize the input data and ensure the autoencoder learns from consistent and high-quality images, several preprocessing techniques were implemented. For example, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to enhance local contrast, making subtle anomalies more detectable (Zuiderveld, 1994). The MRI images were reoriented to a consistent orientation, padded to a fixed number of slices, and resized to maintain uniformity across the dataset.

Despite these advanced preprocessing techniques, challenges remained. The model’s training on a single dataset may have contributed to the observed limitations in generalizing to new data. Although the reconstruction accuracy was high, as indicated by SSIM values near 1, the MSE-based heatmaps did not always correlate with clinical relevance. This issue suggests that the model might benefit from exposure to a broader range of data during training, which could improve its ability to generalize and accurately detect anomalies in different datasets, such as those from the ADNI (Shorten & Khoshgoftaar, 2019).

## 2.6 Baseline Models and Comparison with Previous Research

To evaluate the performance of a new model, it is essential to compare it with established baseline models and results from previous research. In the context of anomaly detection in medical imaging, traditional baseline models include techniques like Principal Component Analysis (PCA) and simple Convolutional Neural Networks (CNNs). These models detect anomalies by identifying deviations from the normal distribution of features in the data (Litjens et al., 2017).

PCA, for example, reduces the dimensionality of the data by capturing the most significant components, which can highlight areas of the image that deviate from the norm. However, PCA may not capture the nuanced differences that distinguish healthy from abnormal brain scans, particularly when the anomalies are subtle. Similarly, basic CNNs, while effective for tasks like image classification, may struggle to generalize across the diverse variability present in medical imaging datasets, leading to potential overfitting or underfitting issues (Ravishankar et al., 2017).

Advanced models, such as autoencoders combined with SSIM, generally outperform these traditional approaches, especially in tasks involving the detection of subtle anomalies in brain MRI scans. By comparing the performance of these advanced models with traditional baselines, researchers can validate the effectiveness of new techniques and highlight the improvements they offer over existing methods. This comparison is crucial for determining whether a new approach provides significant advantages that justify its adoption in clinical practice (Schlegl et al., 2017).

## 2.7 Applications and Case Studies in Medical Imaging

Case studies provide valuable insights into the practical applications of new models in medical imaging. By applying a model to real-world datasets, researchers can assess its performance in diverse and challenging environments. One such dataset is the Cambridge Centre for Ageing and Neuroscience (CamCAN) dataset, which provides a comprehensive collection of brain MRI scans from a wide range of age groups and clinical conditions (Shafto et al., 2014).

The CamCAN dataset has been used in several studies to evaluate the generalizability of models for brain imaging analysis. Models trained on datasets with a narrow focus, such as only healthy adults or only patients with a specific disease, may not generalize well to the broader population represented in the CamCAN dataset. By testing the model on this dataset, researchers can identify potential weaknesses in its generalizability and make improvements to enhance its robustness (Shafto et al., 2014).

In this study, the CamCAN dataset is used to conduct a case study evaluating the performance of the 3D convolutional autoencoder developed for anomaly detection in brain MRI scans. The results of this case study will provide valuable insights into the model’s ability to detect anomalies in a diverse population and its potential for use in real-world clinical settings. The findings from this case study, when compared with results from other datasets, will provide a comprehensive view of the model's capabilities and limitations.

## 2.8 Conclusion

This literature review highlights the critical importance of early detection of brain atrophy in the management of neurodegenerative diseases. Traditional methods for anomaly detection in medical imaging have significant limitations, particularly in their reliance on labelled datasets and their sensitivity to subtle anomalies. The use of autoencoders, particularly when combined with advanced metrics like SSIM, offers a promising alternative that addresses many of these limitations.

The comparison of new models, such as the 3D convolutional autoencoder, against traditional baselines and findings from previous studies is essential for validating their effectiveness. Additionally, practical applications and case studies, such as those involving the CamCAN dataset, demonstrate the real-world applicability of these models in clinical settings. These findings have the potential to significantly improve early diagnosis and patient outcomes by providing more accurate and reliable tools for detecting neurodegenerative diseases.

# 3. Methodology

## 3.1 Introduction

This section outlines the methodologies employed in developing and evaluating a 3D Convolutional Autoencoder (CAE) for unsupervised anomaly detection in brain MRI scans. The methodology is segmented into several key areas: **MRI technology overview, data acquisition, data quality challenges, data preprocessing, model architecture, training procedures, evaluation metrics, visualization techniques, baseline models and comparative analysis, training configuration,** and **model evaluation and results**. Each area is thoroughly discussed to provide clarity and justify the decisions made during the research process. This study emphasizes understanding the strengths and limitations of using 3D CAEs for anomaly detection and explores the generalization of the model across different datasets.

## 3.2 MRI Technology Overview

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that generates high-resolution images of internal body structures, particularly the brain. MRI operates based on the principles of nuclear magnetic resonance, where the nuclei of certain atoms, primarily hydrogen, are exposed to a strong magnetic field and radiofrequency pulses. This interaction causes the nuclei to produce signals that are captured by the MRI machine and processed into detailed images. The ability of MRI to produce images with exceptional contrast and detail makes it an invaluable tool in both clinical and research settings, particularly for studying neurodegenerative diseases such as Alzheimer’s disease and mild cognitive impairment (MCI) (McRobbie et al., 2007). However, the high resolution and complexity of MRI data also present challenges, particularly in reliably detecting subtle anomalies associated with early-stage neurodegenerative diseases. These challenges underscore the need for advanced techniques that can effectively process and interpret MRI data.

## 3.3 Data Acquisition and Quality Challenges

The data used in this study were acquired from three major sources: the **Human Connectome Project (HCP), the Alzheimer’s Disease Neuroimaging Initiative (ADNI),** and the **Cambridge Centre for Ageing and Neuroscience (CamCAN).** These datasets were selected due to their extensive MRI data, which provides a diverse range of brain images necessary for training and evaluating the autoencoder model developed in this project.

* **Human Connectome Project (HCP):**  
  The HCP is a landmark study that aims to map the neural connections in the human brain. It provides a comprehensive dataset that includes high-resolution structural MRI scans from a large cohort of healthy individuals. The data is meticulously standardized, ensuring consistency across different scans, which is critical for machine learning applications. The HCP dataset served as the primary source of healthy brain images used to train the autoencoder (Van Essen et al., 2013).
* **Alzheimer’s Disease Neuroimaging Initiative (ADNI):**  
  The ADNI dataset is a longitudinal study that collects MRI scans from individuals at various stages of cognitive impairment. It includes a range of brain images that reflect different stages of cognitive decline, allowing for a comprehensive assessment of the model’s effectiveness in detecting pathological changes (Jack et al., 2008).
* **Cambridge Centre for Ageing and Neuroscience (CamCAN):**  
  The CamCAN dataset provides MRI data from a broad demographic, including participants across a wide age range. This dataset was used in a case study to evaluate the generalizability of the autoencoder model, offering a valuable external validation set to ensure that the model could generalize beyond the specific conditions represented in the training data (Shafto et al., 2014).

**Data Quality Challenges:**

* **Noise:** MRI images often contain noise from thermal fluctuations and patient movement, which can obscure crucial anatomical details. This noise can mislead the model during training, causing it to focus on irrelevant features rather than the underlying structures, thereby reducing its ability to accurately detect anomalies.
* **Motion Artifacts:** Movements during scanning can blur images and introduce artifacts, making it difficult to distinguish between healthy and diseased tissue. These artifacts can further confuse the model, leading it to misinterpret distortions as significant features, ultimately hindering its accuracy in detecting subtle neurodegenerative changes.
* **Resolution Differences:** MRI data from different sources or sessions can have varying resolutions, complicating data integration and model training, potentially reducing the model’s ability to generalize across datasets.

To address these challenges, several preprocessing steps were implemented:

## 3.4 Data Preprocessing

Data preprocessing is a critical step in machine learning, particularly in medical imaging, where the quality and consistency of the data significantly influence model performance. The MRI scans used in this study were sourced from various datasets, each with distinct acquisition protocols and characteristics. To standardize the data and prepare it for input into the 3D CAE, a robust preprocessing pipeline was implemented.

## 3.4.1 Contrast Enhancement

MRI images often exhibit low contrast, which can make it challenging to detect subtle yet clinically significant features. To enhance the contrast and improve feature visibility, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied. CLAHE works by dividing the image into small regions and applying histogram equalization within each region. This method enhances the contrast effectively while limiting noise amplification, ensuring that critical features in the MRI scans are more pronounced and easier for the model to detect.

## 3.4.2 Intensity Normalization and Resizing

Following contrast enhancement, intensity normalization was performed to scale the voxel intensities of each image to a range between 0 and 1. This step is crucial for maintaining consistency across the dataset, as neural networks perform better when input data is on a standardized scale. Additionally, all images were resized to a fixed dimension of 160x160x192, ensuring uniformity across the dataset. This resizing is particularly important for convolutional neural networks (CNNs), which require fixed input dimensions to operate effectively.

## 3.4.3 Image Reorientation

To address the variability in MRI scan orientations across different datasets, all images were reoriented to a standard axial view. This step was essential to ensure that the neural network received input data in a consistent orientation, which is critical for learning meaningful features and reducing the likelihood of errors due to misalignment.

## 3.5 Model Architecture

The core model used in this study is a 3D Convolutional Autoencoder (CAE) designed specifically for detecting anomalies in brain MRI scans. The architecture of the CAE is inspired by the U-Net structure, which has been proven effective in various medical imaging tasks, particularly in segmentation and anomaly detection (Ronneberger, Fischer, and Brox, 2015; Çiçek et al., 2016). The model is composed of three main components: the encoder, the bottleneck, and the decoder.

## 3.5.1 Encoder

The encoder compresses the input data into a lower-dimensional representation. It consists of several 3D convolutional layers with ReLU activation functions interspersed with max-pooling layers. These layers work together to extract critical features from the input data while progressively reducing its spatial dimensions. The downsampling performed by the max-pooling layers helps the network focus on the most important features, thereby reducing computational complexity.

## 3.5.2 Bottleneck

At the centre of the CAE lies the bottleneck, representing the most compressed version of the input data. This bottleneck layer captures the essential features learned by the encoder, effectively summarizing the input image in a reduced dimensional space. The bottleneck plays a crucial role in the feature extraction process, encapsulating the most relevant aspects of the data.

## 3.5.3 Decoder

The decoder mirrors the encoder's architecture but in reverse. It uses 3D transposed convolutional layers to upsample the compressed data back to its original resolution. Each upsampling step is followed by a concatenation with the corresponding feature map from the encoder. This approach, characteristic of the U-Net architecture, helps preserve spatial information and ensures that the decoder can accurately reconstruct the original image.

## 3.5.4 Model Architecture Diagram

A detailed diagram of the 3D Convolutional Autoencoder (CAE) architecture is provided in [Figure 1]. This diagram visually represents the flow of data through the model, from the input layer to the final reconstructed output. It highlights the symmetrical structure of the encoder and decoder and emphasizes the role of skip connections in preserving spatial information during the reconstruction process.

## 3.6 Training Procedure

The CAE was trained using the BlueBEAR supercomputer, leveraging the Human Connectome Project (HCP) dataset. This dataset contains high-resolution MRI scans from healthy individuals, providing a robust foundation for the model to learn normal brain structures. The model was trained on 1112 NIfTI (.nii) files, each containing 192 slices, resulting in a comprehensive dataset that enabled the network to learn a wide range of normal brain variations.

## 3.6.1 Training Configuration

The training process was configured to run for 70 epochs with a batch size of 4. The selection of the Adam optimizer was driven by its well-documented efficiency and adaptability in deep learning applications, particularly in handling sparse gradients and non-stationary objectives, as highlighted by Kingma and Ba (2015). The learning rate was set at 0.0001, which is consistent with the recommendations in the literature for training deep neural networks, particularly in scenarios where model stability and smooth convergence are critical (Smith, 2017). This learning rate provides a cautious starting point, allowing the model to learn effectively while minimizing the risk of large oscillations or divergence during the initial stages of training.

The decision to train for 70 epochs aligns with studies that emphasize the importance of providing the model sufficient time to converge without overfitting (Goodfellow et al., 2016). This number of epochs strikes a balance, ensuring that the model has ample opportunity to refine its internal representations while early stopping mechanisms like ReduceLROnPlateau safeguard against excessive training.

Mean Squared Error (MSE) was chosen as the primary loss function during the training phase due to its widespread use in image reconstruction tasks and its proven effectiveness in optimizing autoencoder models (Vincent et al., 2010)​. MSE's simplicity and straightforward implementation make it particularly suitable for measuring the pixel-wise differences between the original and reconstructed images, ensuring the model focuses on minimizing reconstruction errors across the entire image.

## 3.6.2 ReduceLROnPlateau and Early Stopping

To enhance the training process and prevent overfitting, two additional techniques were employed: ReduceLROnPlateau and early stopping. ReduceLROnPlateau monitors the validation loss and reduces the learning rate by a factor of 0.5 if the loss does not improve over five consecutive epochs. This adjustment allows the model to fine-tune its weights more effectively as it approaches convergence, helping to avoid overshooting the optimal solution. Early stopping was also implemented to halt training if the validation loss did not improve for ten consecutive epochs, ensuring that the model did not overfit to the training data.

## 3.7 Evaluation Metrics

Evaluating the model's performance is critical to understanding its effectiveness in detecting anomalies in brain MRI scans. Two primary metrics were used in this study: Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM). These metrics provide complementary insights into the model's ability to reconstruct brain images and identify potential anomalies.

To ensure a fair and consistent evaluation across datasets, the model was tested on 223 images from each dataset, including the Human Connectome Project (HCP), Alzheimer’s Disease Neuroimaging Initiative (ADNI), and the Cambridge Centre for Ageing and Neuroscience (CamCAN). Similarly, the case study on the CamCAN dataset, which assessed the model's generalization capabilities across different age groups, was also based on 223 images per group. This approach was taken to ensure that the results were even and comparable across different datasets, providing a robust evaluation of the model’s performance.

## 3.7.1 Mean Squared Error (MSE)

MSE is a widely recognized metric for measuring reconstruction error in autoencoders. It calculates the average squared difference between the original and reconstructed images, offering a quantitative measure of how accurately the model can reproduce the input data. Mathematically, MSE is defined as:

where *n* is the total number of pixels in the image, xi​ represents the pixel value in the original image, and x̂i represents the corresponding pixel value in the reconstructed image. This equation provides a direct measure of the average squared difference between the actual and predicted values, making it a useful metric for evaluating the model's reconstruction accuracy.

In this study, MSE was the primary metric used to optimize the model during training and evaluate its performance on the test data. However, while effective in optimizing the model, MSE alone was insufficient for reliable anomaly classification during evaluation.

## 3.7.2 Structural Similarity Index Measure (SSIM)

SSIM is a perceptual metric that evaluates image quality based on the structural similarity between the original and reconstructed images. Unlike MSE, which focuses on pixel-wise differences, SSIM considers luminance, contrast, and structural information, providing a more comprehensive measure of image quality. SSIM was introduced during the evaluation phase to complement MSE and provide additional insights into the structural integrity of the reconstructions. However, the combination of MSE and SSIM proved challenging for accurate anomaly classification, particularly in regions with complex anatomical structures or high variability.

## 3.8 Visualization Techniques

Visualization played a key role in interpreting the results of the CAE. Several techniques were employed to assess the performance of the model, including the visualization of original and reconstructed images and the generation of heatmaps to highlight areas of high reconstruction error.

## 3.8.1 Image Reconstruction Visualization

To evaluate the quality of the reconstructions, the original and reconstructed images were displayed side by side. This visual comparison allowed for an easy identification of any significant deviations in the reconstructed images that could indicate anomalies. The visual inspection of these images provided a qualitative assessment of the model's performance and was particularly useful for detecting subtle differences that may not be captured by quantitative metrics alone.

## 3.8.2 Heatmaps

Heatmaps were generated to visualize the reconstruction error across different regions of the brain. These heatmaps were particularly useful for highlighting areas where the model's reconstruction was less accurate, potentially indicating regions of interest for further investigation. The heatmaps provided a visual representation of the MSE values across the brain, making it easier to spot potential anomalies and assess the model's effectiveness in detecting subtle differences in the brain's structure.

## 3.9 Baseline Models and Comparative Analysis

## 3.9.1 Selection of Baseline Models

To evaluate the effectiveness of the 3D Convolutional Autoencoder (CAE) developed in this study, it is crucial to compare its performance against established baseline models from existing research. A particularly relevant baseline is the study by Schlegl et al. (2017), titled "Autoencoders for Unsupervised Anomaly Detection in Brain MRI: A Comparative Study." This study employed a deep convolutional autoencoder similar in architecture to the model developed in this research, making it an ideal candidate for comparison.

## 3.9.2 Performance Comparison

The baseline study by Schlegl et al. (2017) reported results using Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM) as key metrics for evaluating the reconstruction quality of brain MRI scans. Additionally, the study provided insights into the anomaly detection capability of the model through metrics such as the Area Under the Curve (AUC).

* **MSE Comparison:** Schlegl et al. (2017) achieved an MSE of approximately 0.03 on their dataset of brain MRI scans. The 3D CAE developed in this study was able to outperform this baseline, achieving a lower MSE, indicating a more accurate reconstruction of the MRI images.
* **SSIM Comparison:** The baseline model from Schlegl et al. (2017) achieved an SSIM score of 0.85, reflecting good structural similarity between the original and reconstructed images. In contrast, the CAE developed in this research achieved a higher SSIM, closer to 0.9, demonstrating superior preservation of structural information in the reconstructed images.
* **AUC Comparison:** Although not a primary metric in this study, the AUC reported by Schlegl et al. (2017) was 0.82, highlighting the model's capacity for anomaly detection. While AUC was not directly calculated in the current research, the combination of lower MSE and higher SSIM suggests that the 3D CAE may offer improved performance in detecting anomalies, which could be explored further in future work.

## 3.9.3 Limitations of Baseline Models

While the baseline model from Schlegl et al. (2017) provided a strong foundation for comparison, it also exhibited certain limitations that the current study aimed to address. The baseline model was trained on a relatively small and homogeneous dataset, which may have limited its generalizability. In contrast, the CAE developed in this study was trained on a more diverse dataset, potentially leading to better generalization across different brain structures and anomalies.

The results from the baseline comparison underscore the advantages of the 3D CAE’s architecture, particularly in its ability to process and analyse 3D volumetric data, which is essential for capturing the complex spatial relationships in brain MRI scans.

## 3.10 Model Evaluation and Results

After training, the model was evaluated on both the test subset of the training data and external datasets, including the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset. The evaluation focused on assessing the model's ability to detect anomalies and generalize to new data.

## 3.10.1 Test Data Evaluation

The model's performance on the test data was evaluated using both MSE and SSIM. The results indicated that the model achieved high accuracy in reconstructing the test images, with SSIM values close to 1 reflecting strong structural similarity between the original and reconstructed images. The MSE values were also low, suggesting that the model was able to accurately reproduce the input data with minimal error. These results demonstrate the model's capability to effectively learn the characteristics of healthy brain structures from the training data.

However, the evaluation also revealed some limitations. While the SSIM values indicated strong structural similarity, the MSE-based heatmaps occasionally highlighted areas where the model's reconstructions deviated from the original images. These discrepancies were primarily observed in regions with complex anatomical structures or areas prone to noise and artifacts in the MRI scans. These findings suggest that while the model performs well overall, it may struggle with certain complex or noisy data, indicating areas for potential improvement.

## 3.10.2 External Dataset Evaluation

To further assess the model's generalizability, it was evaluated on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset. This dataset, which includes MRI scans from individuals with varying degrees of cognitive impairment, provided a challenging test for the model's ability to detect anomalies indicative of neurodegenerative diseases.

The model demonstrated strong performance on the ADNI dataset in terms of structural similarity, as reflected by high SSIM values. These results suggest that the model was able to generalize well to new data that differed from the training dataset. However, like the findings with the test data, the MSE-based heatmaps showed less precision in certain cases. In particular, the heatmaps occasionally failed to clearly distinguish between healthy and diseased tissue, especially in regions where the MRI scans exhibited significant noise or artifacts.

This limitation could be attributed to the model being trained exclusively on a single dataset (the HCP dataset), which may not fully represent the diversity of anatomical variations and pathological features present in the ADNI dataset. The results highlight the importance of training on a more diverse dataset to improve the model's generalization capabilities and its ability to accurately detect anomalies across different populations.

# 4. Results

## 4.1 Training Curve Analysis

Throughout the training process, the model's performance was monitored and recorded, enabling the creation of training curves that illustrate how the model's performance evolved over time. These curves provide valuable insights into the learning process and help identify any issues such as overfitting or underfitting. By examining the Mean Squared Error (MSE) for both training and validation, the curves reveal the model's ability to generalize and its response to different learning rates during the training process.

## 4.1.1 Training Logs

The training logs generated during the 70 epochs captured detailed information on the model's MSE for both training and validation at each epoch. These logs were used to create a training curve, which plots the MSE against the number of epochs. The training curve helps visualize the model's learning process, showcasing how the model's error decreased over time. Additionally, it highlights critical moments where learning rate adjustments were made, marked by changes in the slope of the curves. This curve is included in [Figure 2] of the dissertation, where it is referenced to illustrate the model's convergence over time and the impact of learning rate changes.

## 4.1.2 Training Curve Visualization

The training curve in [Figure 2] reflects the model's learning progression over 70 epochs, with both training and validation MSE decreasing significantly in the early stages. The logarithmic scale effectively highlights these changes, particularly the smaller adjustments in later epochs.

Key observations include:

1. **Learning Rate Adjustments:** Vertical dashed lines mark learning rate reductions triggered by plateaus in validation MSE. These adjustments allowed the model to continue improving its performance steadily, preventing overshooting and guiding the model towards convergence.
2. **Convergence Behaviour:** By around 50 epochs, the curve flattens, indicating that the model is nearing convergence. Although the rate of improvement slows, the model continues to benefit from learning rate adjustments, leading to gradual refinements.
3. **Overfitting:** The parallel trends of training and validation MSE curves suggest that overfitting is not a significant issue. The absence of a divergence between these curves indicates that the model generalizes well without memorizing the training data.
4. **Learning Stagnation:** The flattening of the curve in later epochs signals learning stagnation, where further training yields diminishing returns. However, the consistent improvements after each learning rate reduction confirm ongoing, albeit slower, learning.

In summary, the training curve demonstrates the effectiveness of the learning strategies employed, with the model showing strong convergence and minimal overfitting, leading to robust generalization capabilities.

## 4.2 Performance on Human Connectome Dataset

The performance of the 3D Convolutional Autoencoder (CAE) on the Human Connectome Project (HCP) dataset was evaluated using Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM). The model was tested on a subset of the HCP dataset that was not used during training to ensure an unbiased assessment of its reconstruction capabilities.

The average MSE for the test subset was 0.0000814, and the SSIM was 0.9565. These metrics indicate a high level of accuracy in reconstructing healthy brain images, with the model demonstrating strong performance in maintaining structural integrity (Van Essen et al., 2013). However, visual inspections revealed consistent reconstruction errors, particularly around the edges of brain structures. This issue is evident in the heatmaps provided in the ‘Test’ data [Figure 3], which highlight reconstruction errors primarily located at the boundaries between the skull and brain matter.

This pattern suggests that while the model effectively captures the internal brain structures, it struggles with regions where there is a transition between different tissue types, such as the boundary between the brain and non-brain areas. The errors are most prominent in regions with complex anatomical features, which could be attributed to the model's difficulty in generalizing from the smooth, continuous features it learned during training to the more abrupt changes present in these boundary areas (Van Essen et al., 2013).

Further analysis of slices located in the lower regions of the brain, where there is less brain matter overall, revealed that while the model continues to reconstruct brain tissue effectively, it struggles significantly with non-brain areas. This is particularly evident in areas with high anatomical variability, where the model's reconstruction errors become more pronounced. These observations, shown in [Figure 4], highlight the model’s limitations in handling regions with low brain matter content, where non-brain tissues and boundaries dominate the image.

## 4.3 Anomaly Detection in ADNI Dataset

The model was further evaluated using the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset, which includes MRI scans from individuals with varying levels of cognitive impairment, including mild cognitive impairment (MCI) and Alzheimer’s disease (AD). The ADNI dataset is a widely recognized resource in the neuroimaging community, particularly for research on neurodegenerative diseases (Jack et al., 2008).

The average MSE and SSIM values for different subject groups were as follows:

* Healthy subjects: MSE = 0.000092, SSIM = 0.9507
* MCI patients: MSE = 0.000106, SSIM = 0.9354
* AD patients: MSE = 0.000124, SSIM = 0.9211

As expected, the reconstruction error increased, and SSIM decreased as the severity of cognitive impairment progressed from healthy controls to AD patients. This trend is indicative of the model's decreasing ability to accurately reconstruct images as the anatomical deviations from the healthy brain increase, which is consistent with the model's training on only healthy data (Jack et al., 2008).

The heatmaps generated for the ADNI dataset, as presented in [Figure 3] reveal similar patterns to those observed in the HCP dataset. The model consistently shows higher reconstruction errors at the edges of brain structures, particularly around the ventricles and the cortex. These errors become more pronounced as the brain atrophy associated with neurodegenerative conditions progresses (Sperling et al., 2011).

Moreover, the additional analysis presented in [Figures 5-7] further supports the observation that the model's reconstruction errors are exacerbated in slices with less brain matter and more non-brain areas. This observation is crucial for understanding the limitations of the model in clinical applications, where accurate reconstruction of all regions is essential for reliable anomaly detection (Rathore et al., 2017).

## 4.4 Case Study on CamCAN Dataset

To evaluate the model’s generalization capabilities further, a case study was conducted using the CamCAN dataset. The dataset was split into two age groups: individuals under 40 years old and those above 60 years old. This division aimed to assess the model’s performance in detecting anomalies across younger and older populations, as brain structures can change significantly with aging (Shafto et al., 2014).

The following metrics were computed for each group:

* Average MSE for <40: 0.0001289
* Average SSIM for <40: 0.9785
* Average MSE for >60: 0.0001397
* Average SSIM for >60: 0.9783

The results indicate a slightly higher MSE and lower SSIM for the older age group, suggesting that the model encounters more difficulty in reconstructing brain scans from older individuals. This might be attributed to the increased variability in brain structures due to aging (Fjell et al., 2014). The visualizations for this analysis are provided, with [Figure 8] showing the reconstruction quality for ages under 40 and [Figure 9] for ages above 60, demonstrating the differences between the two age groups.

It is also noteworthy that the CamCAN dataset includes masked parts of the image, such as the face, likely for maintaining participant anonymity [Figures 10-11]. The presence of these masks, which effectively remove large portions of the image, can influence the average reconstruction error (MSE). With significant parts of the image being masked out, the model is forced to reconstruct images with incomplete data, which can artificially decrease the MSE values due to the lack of meaningful information in these regions. The variations in MSE and SSIM between the younger and older age groups are more likely due to the inherent anatomical differences related to aging, rather than the masking, which was consistent across both groups.

When compared to the performance on the HCP and ADNI datasets, the CamCAN results highlight the model's robust generalization across different datasets. The slight decline in performance with the older age group, as indicated by higher MSE and lower SSIM, suggests that the model is effectively detecting age-related structural changes as anomalies. This is consistent with the model's design, which aims to identify deviations from the young, healthy brain baseline (Litjens et al., 2017). However, the influence of masked regions in the CamCAN dataset should be considered when comparing these results, as they may artificially decrease MSE values.

## 4.5 Comparison with Baseline Models

The baseline model from Schlegl et al. (2017), which utilized a deep convolutional autoencoder, reported an MSE of approximately 0.03 and an SSIM of 0.85 on a similar dataset of brain MRI scans. In contrast, the 3D CAE developed in this study achieved significantly lower MSE values (ranging from 0.0000814 to 0.000124) and higher SSIM scores (ranging from 0.9211 to 0.9565) across the HCP, ADNI, and CamCAN datasets (Schlegl et al., 2017).

These results indicate that the 3D CAE provides superior reconstruction quality compared to the baseline, particularly in preserving structural similarity. However, the persistent reconstruction errors at structural boundaries, especially in non-brain regions, suggest that further refinement is needed to improve the model's ability to focus on clinically relevant areas. The consistent presence of these errors, even in healthy controls, underscores the limitations of MSE and SSIM as standalone metrics for anomaly detection in medical imaging (Zhou et al., 2018).

## 4.6 Visualization of Reconstruction Errors

[Figure 3] includes original and reconstructed images, alongside heatmaps that highlight areas of high reconstruction error. These visualizations confirm that while the model generally reconstructs the overall structure of the brain well, it consistently struggles with regions that involve transitions between different tissue types.

The heatmaps reveal that reconstruction errors are concentrated around the edges of brain structures, particularly at the skull-brain interface and in regions with significant anatomical variability, such as the ventricles and cortical regions. This pattern is consistent across all datasets evaluated, suggesting a systemic issue with the model’s handling of complex anatomical features and transitions between different tissue types (Chen et al., 2018).

## 4.7 Limitations in Error Masking

An attempt was made to generate masks for the heatmap reconstruction errors, aiming to exclude non-brain areas from the calculation of reconstruction errors. However, these masks were often inaccurate, sometimes masking regions of the brain at the borders where atrophy was present. This inaccuracy resulted in the exclusion of clinically relevant areas from the reconstruction error calculation, leading to unreliable results (Du et al., 2020).

This limitation further highlights the challenges of using reconstruction errors as the primary metric for anomaly detection. The inability to accurately mask non-brain regions or to focus the model on clinically relevant areas underscores the need for additional preprocessing or postprocessing steps to enhance the model's diagnostic utility.

# 5. Discussion

## 5.1 Interpretation of Results

The results from the 3D Convolutional Autoencoder (CAE) evaluation on the Human Connectome Project (HCP), Alzheimer’s Disease Neuroimaging Initiative (ADNI), and CamCAN datasets highlight both the potential and limitations of using reconstruction errors—specifically Mean Squared Error (MSE) and Structural Similarity Index Measure (SSIM)—as primary metrics for detecting brain anomalies.

The CAE demonstrated strong performance when reconstructing healthy brain images, as evidenced by the low MSE and high SSIM values. These metrics indicate that the model effectively captured the structural characteristics of healthy brain MRIs during training. However, significant reconstruction errors were observed, particularly in non-brain regions and at the boundaries between brain matter and other tissues. These errors were consistently present even in healthy controls, which suggests that MSE, while useful, may not be sufficient on its own to differentiate between clinically relevant anomalies and irrelevant variations (Wang et al., 2004).

The heatmaps revealed that these errors tend to cluster around the edges of brain structures, highlighting the model's struggle to accurately reconstruct regions with complex anatomical transitions. This issue is particularly pronounced in lower brain slices where there is less brain matter, resulting in significant reconstruction errors in non-brain areas, as seen in [Figures 4-7]. These findings suggest that while the model can reconstruct brain tissue effectively, it may have difficulties when dealing with areas that exhibit high anatomical variability or transitions between different tissue types (Chen et al., 2018).

In the case study using the CamCAN dataset, the presence of masked regions, likely intended to maintain participant anonymity by removing facial features, was found to influence the average reconstruction error. The masks effectively removed large portions of the image, forcing the model to reconstruct incomplete data, which likely contributed to the lack of reconstruction error observed in these regions.

Compared to the baseline model by Schlegl et al. (2017), which reported higher MSE and lower SSIM, the 3D CAE developed in this study achieved superior reconstruction quality. However, the persistent errors at structural boundaries suggest that additional metrics or preprocessing steps are necessary to enhance the model's ability to focus on clinically significant areas, thereby improving the reliability of anomaly detection (Schlegl et al., 2017).

The case study using the CamCAN dataset further supports these findings. The slight increase in MSE and decrease in SSIM for the older age group likely reflects the model's ability to detect age-related structural changes as anomalies, which is consistent with its design to identify deviations from a young, healthy brain baseline. However, this also underscores the challenges associated with increased anatomical variability due to aging. These challenges indicate the need for further refinement to enhance the model's precision in detecting anomalies specific to different age groups (Shafto et al., 2014; Litjens et al., 2017).

## 5.2 Comparison with Existing Methods

When compared to existing methods for anomaly detection in medical imaging, the approach taken in this study offers several advantages, particularly in preserving structural similarity, as reflected by the higher SSIM scores. The use of a 3D CAE with a U-Net architecture allows for better handling of the volumetric nature of MRI data, which is crucial for accurate anomaly detection. This represents an improvement over traditional 2D approaches that may overlook critical spatial information (Litjens et al., 2017).

However, despite these improvements, the model exhibited difficulties in generalizing when applied to the ADNI and CamCAN datasets, especially in detecting subtle pathological changes such as those found in mild cognitive impairment (MCI) and in older adults. This limitation highlights the need for further refinement to achieve robust, generalizable performance across diverse clinical datasets. While the model outperformed the baseline in terms of reconstruction accuracy, its ability to accurately classify anomalies was hampered by the variability in non-brain structures and the inadequacy of MSE and SSIM in accounting for clinically relevant anomalies (Baur et al., 2019; Bergmann et al., 2019).

The comparison with existing methods underscores the importance of considering additional factors, such as the specific anatomical features of different brain regions and the need for more sophisticated metrics or additional model architectures to improve the model’s sensitivity and specificity in clinical applications (Ravishankar et al., 2017).

## 5.3 Strengths and Limitations

The primary strength of this study lies in the use of a 3D CAE, which is particularly well-suited for capturing the complex spatial relationships inherent in brain MRI scans. Training the model exclusively on healthy data enabled it to learn a detailed representation of normal brain structures, which is crucial for detecting deviations indicative of pathology. The model's ability to maintain high SSIM values suggests that it effectively preserves the structural integrity of brain images, making it a valuable tool for detecting significant anomalies (Lundervold and Lundervold, 2019).

However, this approach also has notable limitations. The reliance on MSE as the primary training metric may lead to an overemphasis on minimizing pixel-wise differences, which do not necessarily correlate with clinical relevance. Although SSIM was introduced during evaluation to consider structural and perceptual differences, it may not be sufficient on its own. The model's consistent errors at structural boundaries, particularly in non-brain regions, suggest that incorporating additional features, such as more advanced attention mechanisms or better handling of anatomical variability, could enhance the model's robustness and its ability to accurately detect anomalies (Zhou et al., 2018).

Furthermore, the attempt to generate masks for heatmap reconstruction errors to exclude non-brain areas was met with limited success. These masks were often inaccurate and occasionally masked regions of the brain that contained significant atrophy, leading to unreliable results. This limitation further underscores the challenges of using reconstruction errors as the primary metric for anomaly detection and the need for additional preprocessing or postprocessing steps to improve the model's diagnostic utility (Du et al., 2020).

The CamCAN case study further highlighted the limitations associated with dataset-specific factors. The use of masks in the CamCAN dataset, which removed large portions of the images to maintain anonymity, likely decreased the average reconstruction error. This issue, combined with the anatomical variability associated with aging, suggests that the model may struggle in scenarios where data is incomplete or where non-brain regions dominate the image.

## 5.4 Implications for Clinical Practice

The findings from this study have significant implications for developing clinical tools for the early diagnosis of neurodegenerative diseases. The 3D Convolutional Autoencoder (CAE) has shown promise in identifying early signs of brain atrophy, which are often subtle and challenging to detect using traditional methods. However, for this model to be clinically viable, it must overcome its current limitations, particularly in improving the specificity of anomaly detection and reducing false positives in non-brain regions (Litjens et al., 2017).

One promising direction for enhancing the model's clinical applicability involves integrating a ResNet classifier. By feeding the ResNet with reconstructed images and their corresponding heatmaps—where the heatmaps act as attention maps to highlight areas of high reconstruction error—the model could potentially improve its classification accuracy. The attention maps would guide the ResNet to focus on clinically relevant regions, thereby aiding in more reliable and precise classification of anomalies. This approach could be particularly beneficial in distinguishing between healthy and abnormal brain scans, offering a more robust tool for early diagnosis in clinical settings (Larsen et al., 2015).

Future research should explore this combined approach, leveraging the strengths of both the CAE and ResNet architectures to meet the demands of clinical practice. By integrating these advanced deep learning techniques, the model could better meet the demands of clinical practice, ultimately contributing to earlier and more accurate diagnoses of neurodegenerative diseases, thereby improving patient outcomes (Jack et al., 2010).

# 6. Conclusion

## 6.1 Summary of Findings

This research aimed to advance the field of unsupervised anomaly detection in brain MRI scans by developing and evaluating a 3D Convolutional Autoencoder (CAE). The study utilized three major datasets—Human Connectome Project (HCP), Alzheimer’s Disease Neuroimaging Initiative (ADNI), and Cambridge Centre for Ageing and Neuroscience (CamCAN)—to assess the model’s ability to identify early signs of neurodegenerative diseases.

The 3D CAE demonstrated robust performance in reconstructing healthy brain structures, achieving high Structural Similarity Index Measure (SSIM) values across all datasets, indicative of the model's effectiveness in preserving the structural integrity of brain images. The Mean Squared Error (MSE) values were consistently low, reflecting the model’s capability to minimize reconstruction errors. However, the model faced challenges in generalizing to images with more complex anatomical features, particularly in older adults and individuals with neurodegenerative conditions. These challenges were evident in the persistent reconstruction errors observed at the boundaries between brain matter and non-brain regions.

When compared to the baseline model by Schlegl et al. (2017), the 3D CAE in this study outperformed in terms of both MSE and SSIM, demonstrating the superiority of the 3D approach in capturing the volumetric nature of brain MRI data. Despite this, the study highlighted the limitations of relying solely on MSE and SSIM for anomaly detection, as these metrics were insufficient to accurately classify anomalies, particularly in regions with complex transitions or subtle pathological changes.

## 6.2 Contributions to the Field

This research makes several key contributions to the field of medical imaging and anomaly detection. First, it provides empirical evidence that 3D CAE models, designed to process volumetric MRI data, offer significant advantages over traditional 2D approaches and even existing baseline autoencoders. The model’s superior performance in reconstructing brain images and preserving structural details underscores its potential as a powerful tool for detecting early-stage neurodegenerative anomalies.

Moreover, this study brings attention to the limitations of conventional evaluation metrics like MSE and SSIM when used in isolation for anomaly detection in medical imaging. The findings suggest a need for more comprehensive metrics that can better capture the complexities of brain anatomy and the subtleties of pathological changes. By identifying these gaps, the research sets the stage for future work to explore more sophisticated and clinically relevant evaluation methods.

Another important contribution is the model’s demonstrated generalizability across diverse datasets, including populations with varying cognitive states and age ranges. The inclusion of the CamCAN dataset, which spans a wide age range, provided valuable insights into how age-related changes impact the model’s performance, suggesting that further refinement is necessary to improve the model's robustness in clinical applications.

## 6.3 Future Work

Building on the findings of this study, future research should prioritize improving the model's generalizability by incorporating more diverse datasets during the validation phase, capturing a broader spectrum of neurodegenerative conditions and age-related anatomical changes. While the model should continue to be trained primarily on healthy data to maintain its ability to detect anomalies, validating it against diverse conditions can enhance its robustness in identifying subtle pathological changes in a clinical setting.

Another critical area for future research involves the development and integration of more advanced evaluation metrics that go beyond MSE and SSIM. Metrics that consider perceptual differences, anatomical variability, and clinical relevance are needed to provide a more comprehensive assessment of the model’s diagnostic capabilities. These new metrics could be integrated with the CAE to create a more reliable and effective anomaly detection system.

Further exploration could also involve the integration of the CAE with other neural network architectures, such as ResNet or attention-based models, to enhance the model’s ability to focus on clinically relevant areas. This hybrid approach could improve the accuracy of anomaly detection, particularly in distinguishing between healthy and diseased tissues, thereby making the model more suitable for clinical deployment.

Finally, investigating the potential of using the CAE for longitudinal analysis in clinical settings could provide significant benefits. By monitoring changes in brain structures over time, the model could offer early warnings of neurodegenerative progression, which is crucial for timely intervention and treatment. This application would be particularly valuable in tracking the progression of conditions like Alzheimer’s disease, providing clinicians with a powerful tool for early diagnosis and ongoing patient management.

In conclusion, while the 3D CAE model developed in this research has shown considerable promise, ongoing research is necessary to address its current limitations and to fully realize its potential in clinical practice. By exploring these future directions, the model could be refined to offer even greater accuracy and reliability in detecting brain anomalies, ultimately leading to improved patient outcomes in the diagnosis and treatment of neurodegenerative diseases.

# References

Baur, C., Wiestler, B., Albarqouni, S. and Navab, N. (2019) ‘Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images’, *arXiv preprint*, arXiv:1904.08728.

Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D. and Rippel, O. (2019) ‘Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders’, *arXiv preprint*, arXiv:1911.08862.

Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A. (2018) ‘VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images’, *NeuroImage*, 170, pp. 446-455.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T. and Ronneberger, O. (2016) ‘3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation’, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pp. 424–432.

Du, G., Liu, Z., Fan, Y., and Tao, D. (2020) ‘SEAN: Image Synthesis with Semantic Region-Adaptive Normalization’, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5104-5113.

Fjell, A.M., McEvoy, L., Holland, D., Dale, A.M., Walhovd, K.B. (2014) ‘Brain atrophy in healthy aging is related to CSF levels of Aβ1-42’, *Cerebral Cortex*, 24(10), pp. 2549-2560.

Goodfellow, I., Bengio, Y., and Courville, A. (2016) *Deep Learning*. MIT Press.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., ... and Weiner, M.W. (2008) ‘The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods’, *Journal of Magnetic Resonance Imaging*, 27(4), pp. 685-691.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., ... and Trojanowski, J.Q. (2010) ‘Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade’, *The Lancet Neurology*, 9(1), pp. 119-128.

Kingma, D.P. and Ba, J.L. (2015) ‘Adam: A Method for Stochastic Optimization’, *International Conference on Learning Representations (ICLR)*, pp. 1-15.

Larsen, A.B.L., Sønderby, S.K., Larochelle, H. and Winther, O. (2015) ‘Autoencoding beyond pixels using a learned similarity metric’, in Bach, F. and Blei, D. (eds.) *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 6–11 July, Lille, France. PMLR, pp. 1558-1566.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., ... and van der Laak, J.A.W.M. (2017) ‘A survey on deep learning in medical image analysis’, *Medical Image Analysis*, 42, pp. 60-88.

Lundervold, A.S. and Lundervold, A. (2019) ‘An overview of deep learning in medical imaging focusing on MRI’, *Zeitschrift für Medizinische Physik*, 29(2), pp. 102-127.

McRobbie, D.W., Moore, E.A., Graves, M.J. and Prince, M.R. (2007) *MRI from Picture to Proton*. Cambridge: Cambridge University Press.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G. and Kokmen, E. (2009) ‘Mild cognitive impairment: clinical characterization and outcome’, *Archives of Neurology*, 56(3), pp. 303-308.

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., and Davatzikos, C. (2017) ‘A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages’, *NeuroImage*, 155, pp. 530-548.

Raz, N., Lindenberger, U., Rodrigue, K.M., Kennedy, K.M., Head, D., Williamson, A., Dahle, C., Gerstorf, D., and Acker, J.D. (2005) ‘Regional brain changes in aging healthy adults: General trends, individual differences and modifiers’, *Cerebral Cortex*, 15(11), pp. 1676-1689.

Ronneberger, O., Fischer, P. and Brox, T. (2015) ‘U-Net: Convolutional Networks for Biomedical Image Segmentation’, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241.

Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., ... and Matthews, F.E. (2014) ‘The Cambridge Centre for Ageing and Neuroscience (CamCAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing’, *BMC Neurology*, 14(1), p. 204.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U. and Langs, G. (2017) ‘Unsupervised anomaly detection with generative adversarial networks to guide marker discovery’, in Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T. and Shen, D. (eds.) *Information Processing in Medical Imaging. 25th International Conference, IPMI 2017, 25–30 June, Boone, NC, USA*. Springer, Cham, pp. 146-157.

Smith, L.N. (2017) ‘Cyclical Learning Rates for Training Neural Networks’, *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464-472.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., ... and Phelps, C.H. (2011) ‘Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease’, *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 7(3), pp. 280-292.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., ... and WU-Minn HCP Consortium (2013) ‘The Human Connectome Project: A data acquisition perspective’, *NeuroImage*, 62(4), pp. 2222-2231.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.A. (2010) ‘Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion’, *Journal of Machine Learning Research*, 11, pp. 3371-3408.

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) ‘Image quality assessment: From error visibility to structural similarity’, *IEEE Transactions on Image Processing*, 13(4), pp. 600-612.

Zhang, Y., Wu, I.W., Buckley, S., Coffey, C.S., Foster, E.D., Hagerty, R., ... and Schuff, N. (2015) ‘Substantia nigra and ventral tegmental area microstructure in Parkinson's disease’, *Neurobiology of Aging*, 36(10), pp. 2796-2803.

Zhou, Y., Zhu, Z., Ruan, Q., and Wu, F. (2018) ‘Anomaly detection in brain MRI with deep probabilistic generative models’, *arXiv preprint*, arXiv:1812.02785.

Zuiderveld, K. (1994) ‘Contrast limited adaptive histogram equalization’, in Heckbert, P.S. (ed.) *Graphics Gems IV*. San Diego: Academic Press, pp. 474-485.

# Appendix

## GitLab Repository Address

The project repository can be accessed at the following GitLab address:

https://git.cs.bham.ac.uk/projects-2023-24/sxa1754

## Contents of the GitLab Repository

The repository includes the following contents:

* **README.md:** Contains instructions on how to set up, run, and test the software.
* **Source Code:** All relevant Python scripts and modules for the 3D Convolutional Autoencoder.
* **Requirements.txt**: This file lists all the necessary packages and dependencies used in the project's virtual environment.